LE TAL ET LES MOTS

I INTRODUCTION

Quelles sont les opérations sur les mots que l'on cherche à faire automatiquement ?

1) Etiqueter les mots (tagueurs)

2) Découper les mots (analyse morphologique)
PC-Kimmo : http://www.sil.org/pckimmo

3) Classer les mots (analyse sémique ou sémantique lexicale, dictionnaire électronique...)
Wordnet : http://www.cogsci.princeton.ed/~wn

4) Corriger les mots (Cf TD 2)

5) Retrouver des mots dans un texte ou plus largement dans un ensemble de documents (concordances, indexation automatique, recherche ou extraction d'informations... CF CM 3)

...

Objectifs visés, résultats atteints, données construites (ie les ressources) et méthodes utilisées.

II L'ETIQUETAGE

• Objectif :
passer d'un texte brut exempt d'informations linguistiques, à un texte dit étiqueté, ie un texte enrichi d'informations diverses.

• Plusieurs opérations de niveau inférieur
- la délimitation des mots

- choix d'une liste d'étiquettes pertinentes

- levée des ambiguïtés lexicales

(1a) *Le vol partira à onze heure. Air France s'y engage.*
(1b) *Le vol a été commis à onze heure. La police en a la preuve.*
(2) *La belle ferme la maison*

Le contexte permet de savoir quelle étiquette sélectionner. Contexte (syntaxique) ou cotexte ?

• Méthodes utilisées pour étiqueter

- Un dictionnaire et une grammaire de levée d'ambiguïté.

Exemples de règles :
* *det det*
* *verbe conj verbe conj*

- Un corpus de textes étiquetés.

- Un corpus non étiqueté mais accompagné d'informations linguistiques sur la relation étiquettes-suffixes.

Systèmes à base de corpus = apprentissage

Statisques
+ éventuellement
heuristiques...

...qui peuvent "cacher" des connaissances théoriques. Méthodes symboliques mais avec connaissances linguistiques implicites.

III EVALUATION DES RESULTATS

1) Quel est le jeu d'étiquettes choisies ?

- Jeu petit ----> moins d'erreurs possibles
- Jeu très diversifié ----> plus d'ambiguïté morphologique, mais aussi plus de précision.

10 à 500 étiquettes différentes.

2) Quelle est la capacité du système à être améliorié, corrigé, réutilisé pour d'autres applications ?

3) Quelques mesures :

le taux de bruit : la proportion d'étiquettes inexactes parmi les étiquettes présentées en sortie.

le taux de silence : la proportion d'étiquettes non présentées en sortie parmi les étiquettes attendues.

le taux de précision : la proportion d'étiquettes exactes parmi les étiquettes présentées (complémentaire du bruit)

le taux de rappel : la proportion d'étiquettes présentées parmi les étiquettes attendues (complémentaire du silence)

• Les résultats obtenus : 1% de bruit et 30 % de silence (Brill - ftp://ftp.cs.jhu.edu/pub/brill).

IV REDEFINITION DU MOT

1) Difficultés propres au TAL

- L'apostrophe

l'arbre , j'arrive... • *aujourd'hui , d'abord...*

- Le tiret

porte-monnaie, c'est-à-dire,...

• *voulez-vous, arrive-t-il ....*

2) Les frontières de mots ?

- Mot *vs* mot composé

chou-fleur, portefeuille, anthropologue, , socio-culturel
carte grise, chemin de fer, timbre-poste, chaise longue
pomme de terre, cochon de lait, chaise longue

- Mot *vs* terme

3) Le traitement de mots inconnus

- les noms propres
- les néologismes
- les emprunts

camping, adagio
redingote <-- riding coat.

- les abréviations et les sigles

prof, bac, fac, ciné...
prolo, apéro pitaine, bus
SMIC Smicard ENA Enarque

CONCLUSION

Le TAL a très largement contribué à renouveler les problématiques en morphologie .

BIBLIOGRAPHIE

Fuchs, Catherine, 1993,
Linguistique et traitements automatiques des langues, Hachette Université.

Laporte, Eric, 1997,
"Les mots. Un demi-siècle de traitements", in Revue TAL, numéro 38, 1997, *Etat de l'Art.*